Coincidence Component Analysis – CCA

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

25rd Oct 2021

Abstract

Principal component analysis (PCA) is a dimensionality reduction based on an orthogonal projection that maximizes data variance along the first axes. The Jaccard similarity index has also been used for dimensionality reduction in kernel methods assuming non-negative random variables values. The present work addresses the use of the Jaccard and coincidence indices adapted to take into account negative values, and compare the results with respect to the iris dataset. The obtained results indicate that the three approaches lead to markedly similar projections, implying that, at least for the considered data, that the Jaccard and coincidence indices approaches adapted to negative random variable values constitute an interesting alternative to traditional PCA given the smaller computational expenses involved in the calculation of the respective joint variation matrices. In addition, given that the coincidence index imposes a more strict consideration of joint variation, as it also takes into account the interiority index, the results provided by this method could be more representative in specific applications.

'*In a chorus of wind-shaped bushes, a windmill sings.*'

LdaFC

1 Introduction

Joint variation between two random variables has been traditionally quantified in terms of conjoint moments including the covariance and Pearson correlation coefficient. Given a collection of N random variables, or measurements, it is possible to calculate the respective $N \times N$ covariance or Pearson correlation coefficients, followed by the determination of the respective eigenvalues and eigenvectors, which are then used to project the original dataset into a new space such that the maximum data variance is concentrated in the first axes, which gives rise the extensively applied Principal Component Analysis – PCA [1, 2, 3, 4, 5]. As a consequence of this property, it becomes possible to obtain data dimensionality reduction by not considering the axes associated to small variances.

Subsequently, several projection methods were proposed, including the family of kernel-based PCA (e.g. [6]) of which the Jaccard for measurable functions (non-negative random variable values) constitutes a positive definite matrix.

The coincidence product between two vectors or functions consists of a functional analogous to the traditional inner product that, however, combines the Jaccard and interiority indices [7, 8, 9], therefore providing a more strict quantification of the joint variation between two random variables.

The coincidence product has been shown to provide enhanced results when comparing clusters and, when employed in a respective convolution, provide improved narrower and sharper peaks when applied to template matching, also allowing substantial reduction of the secondary matches [10].

In the present work, we present the Jaccard PCA and coincidence component analysis (CCA) derived from the coincidence index adapted to cope with negative random variables values. The respective potential of these approaches, including the traditional PCA, are them compared with respect to the standardized version of the iris dataset, yielding surprisingly similar results.

2 Principal Component Analysis

Let X_i, $i = 1, 2, \ldots, N$, be N random variables associated to respective observations of measurements or properties of N entities. For instance, each entity could be a fruit, and the random variables could correspond to respective properties such as weight, size, etc.

Let each of the N entities i be associated to a respective

feature vector:

$$\vec{X}_i = \begin{bmatrix} X_{i,1} \\ X_{i,2} \\ \cdots \\ X_{i,N} \end{bmatrix} \tag{1}$$

The traditional method called Principal Component Analysis starts by obtaining matrices corresponding to the estimated covariance K or Pearson correlation P coefficients between the N involved random variables.

The eigenvalues $\vec{v}_i$ are then organized into the following transformation matrix:

$$T = \begin{bmatrix} \leftarrow & \vec{v}_1 & \rightarrow \\ \leftarrow & \vec{v}_2 & \rightarrow \\ \cdots & \cdots & \cdots \\ \leftarrow & \vec{v}_N & \rightarrow \end{bmatrix} \tag{2}$$

Each new random variable, obtained through a coordinate system rotation (recall that K or P are non-negative and symmetric), can now be obtained as:

$$\tilde{\vec{X}}_i = T\vec{X}_i \tag{3}$$

3 Jaccard PCA

The Jaccard similarity index (e.g. [11, 12, 7]) has been extensively employed in the most diverse areas for its conceptual simplicity and effectiveness. Given any two sets A and B, the respective Jaccard index can be determined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{4}$$

with $0 \leq J(A,B) \leq 1$.

In the Jaccard PCA (e.g. [6]), which requires all random variables to have non-negative values, the following matrix is used in place of the covariance or Pearson correlation coefficient:

$$C[X,Y] = \frac{\sum_{i=1,N} \min(X_i, Y_i)}{\sum_{i=1,N} \max(X_i, Y_i)} \tag{5}$$

with $0 \leq J(X_i, X_j) \leq 1$.

This result derives from multiset theory (e.g. [13, 14, 15, 16, 17, 18]), by understanding the function values as real-valued versions of the elements multiplicity.

The multiset Jaccard index generalized to negative multiplicities can be adapted [7, 8, 10] for taking into account the negative multiplicities as:

$$\mathcal{J}_N(X,Y) = \frac{\ll X,Y \gg}{X \diamond Y} \tag{6}$$

where:

$$\ll X,Y \gg = \sum_{i=1}^{N} s_{X_i} s_{Y_i} \min(s_{X_i} X_i, s_{Y_i} X_j) \tag{7}$$

and:

$$X \diamond Y = \sum_{i=1}^{N} \max(s_{X_i} X_i, s_{Y_i} X_j) \tag{8}$$

4 Coincidence Component Analysis

The coincidence product between two random variables [7, 10] X and Y can be simply expressed as:

$$\mathcal{C}(X,Y) = \mathcal{J}_N(X,Y)\, \mathcal{I}_N(X,Y) \tag{9}$$

The interiority index adapted to possibly negative multiplicities [10] corresponds to:

$$\mathcal{I}_N(X,Y) = \frac{\ll X,Y \gg_+}{\min\{A_X, A_Y\}} \tag{10}$$

where:

$$\ll X,Y \gg_+ = \sum_{i \in S_+} \min(s_{X_i} X_i, s_{Y_i} X_j) \tag{11}$$

with $S_+ = \{i | s_{X_i} s_{Y_i} > 0\}$.

and:

$$A_X = \sum_{i=1}^{N} |X_i| \tag{12}$$

$$A_Y = \sum_{i=1}^{N} |Y_i| \tag{13}$$

$$\tag{14}$$

The Coincidence Component Analysis – CCA – then consists in employing the coincidence matrix instead of the covariance K or Pearson correlation coefficient P matrices. As such, a respective transformation matrix C is obtained, so that the CCA can now be expressed as the following linear statistical transformation:

$$\hat{\vec{X}}_i = C\vec{X}_i \tag{15}$$

5 A Case Example

Figure 1 illustrates a scatterplot obtained from the two main axes in the PCA, Jaccard PCA and CCA of the iris dataset, considering the respectively indicated combinations of features. All four features, which are originally non-negative, were standardized prior to the application of the three methodologies.

The obtained projections onto two axes reveal a surprising similarity of performance for all three considered methods while considering combinations of 4 and 3 features. There are, however, relatively small differences that are accounted by the different measurements of joint variation giving rise to each of the considered methods.

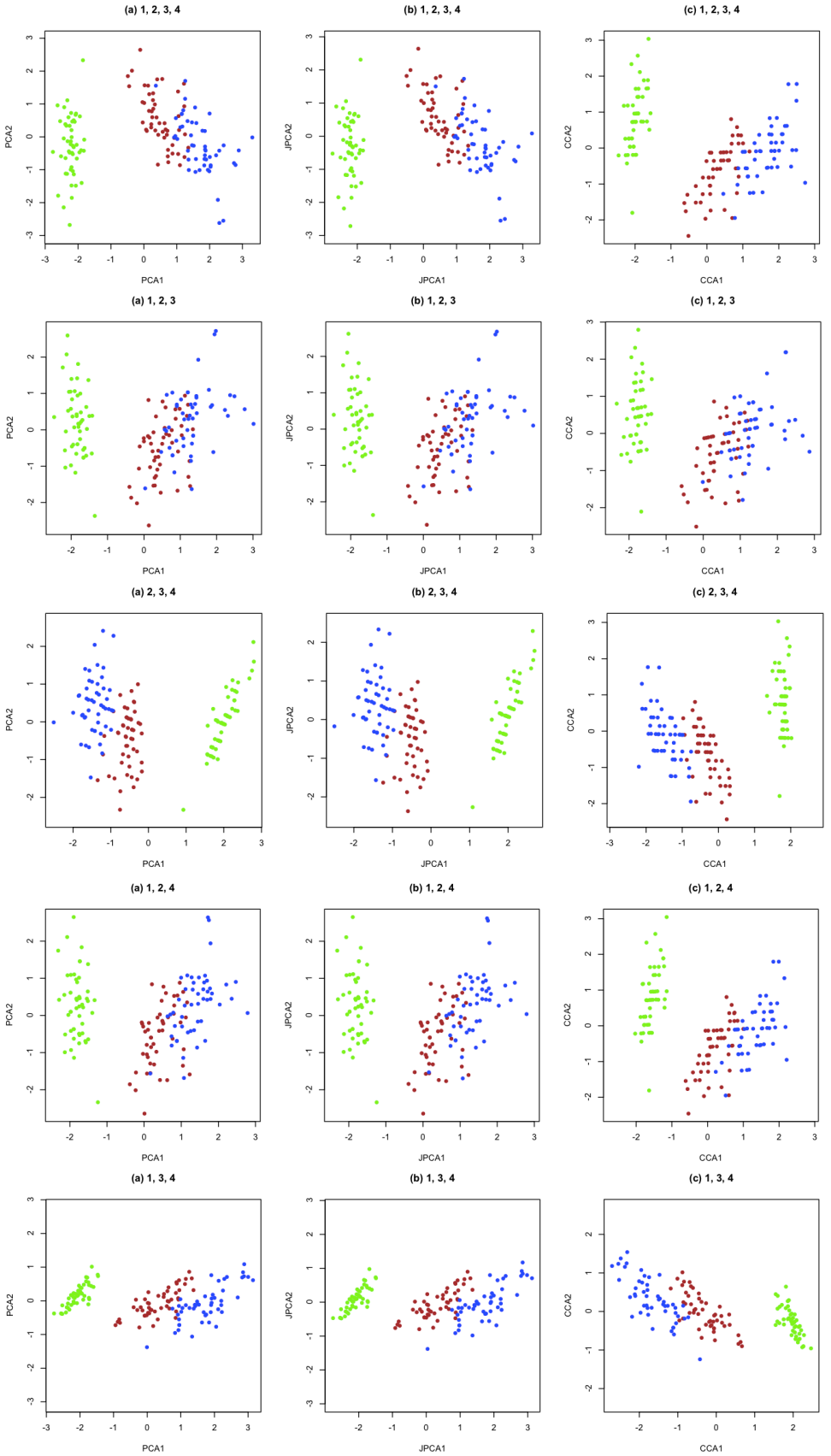

Figure 1: PCA, Jaccard PCA, and Coincidence Component Analysis of the iris dataset for four combinations of the 4 original features. Remarkably similar results have been obtained.

6 Concluding Remarks

The stochastic method known as Principal Component Analysis has been extensively employed in the most diverse areas as an effective means for analysis of statistical data, especially as a means to achieve dimensionality reduction.

In the present work, we presented the extension of the Jaccard PCA to cope with negative random variable values. The new Coincidence Component Analysis methodology was also introduced, in an extended version to cope with negative values.

The adaptations of the Jaccard PCA and CCA to negative values allowed them to be compared one another and also with the traditional PCA. Results regarding the iris dataset indicated that the three methods perform similarly, though relatively small differences can be observed. In the case of the CCA, these differences have to do with the relatively more strict quantification of joint variation implied by the adoption of the coincidence index.

Acknowledgments.

Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) and FAPESP (grant 15/22308-2).

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, 2000.

[2] L. da F. Costa and R. M. C. Cesar Jr. *Shape Classification and Analysis: Theory and Practice*. CRC Press, Boca Raton, 2nd edition, 2009.

[3] K. Koutrombas and S. Theodoridis. *Pattern Recognition*. Academic Press, 2008.

[4] R. A. Johnson and D.W. Wichern. *Applied multivariate analysis*. Prentice Hall, 2002.

[5] F. L. Gewers, Gustavo R. Ferreira, H. F. Arruda, F. N. Silva, C. H. Comin, D. R. Amancio, and L. da F. Costa. Principal component analysis: A natural approach to data exploration. *ACM Computing Surveys*, pages 1–34, 2021.

[6] Wikipedia. Kernel principal component analysis. Wikipedia, the free encyclopedia, 2021. https://en.wikipedia.org/wiki/Kernel_principal_component_analysis. [Online; accessed 10-Apr-2020.].

[7] L. da F. Costa. Further generalizations of the Jaccard index. https://www.researchgate.net/publication/355381945_Further_Generalizations_of_the_Jaccard_Index, 2021. [Online; accessed 21-Aug-2021].

[8] L. da F. Costa. Multisets. https://www.researchgate.net/publication/355437006_Multisets, 2021. [Online; accessed 21-Aug-2021].

[9] L. da F. Costa. Analogies between boolean algebra, set theory, and function spaces. https://www.researchgate.net/publication/355680272_Analogies_Between_Boolean_Algebra_Set_Theory_and_Function_Spaces, 2021. [Online; accessed 21-Aug-2021].

[10] L. da F. Costa. Comparing cross correlation-based similarities. https://www.researchgate.net/publication/355546016_Comparing_Cross_Correlation-Based_Similarities, 2021. [Online; accessed 21-Aug-2021].

[11] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société vaudoise des sciences naturelles*, 37:547–549, 1901.

[12] Wikipedia. Jaccard index. https://en.wikipedia.org/wiki/Jaccard_index. [Online; accessed 10-Oct-2021].

[13] J. Hein. *Discrete Mathematics*. Jones & Bartlett Pub., 2003.

[14] D. E. Knuth. *The Art of Computing*. Addison Wesley, 1998.

[15] W. D. Blizard. Multiset theory. *Notre Dame Journal of Formal Logic*, 30:36—66, 1989.

[16] W. D. Blizard. The development of multiset theory. *Modern Logic*, 4:319–352, 1991.

[17] P. M. Mahalakshmi and P. Thangavelu. Properties of multisets. *International Journal of Innovative Technology and Exploring Engineering*, 8:1–4, 2019.

[18] D. Singh, M. Ibrahim, T. Yohana, and J. N. Singh. Complementation in multiset theory. *International Mathematical Forum*, 38:1877–1884, 2011.